Bridging LLM Reasoning and UMLS for Spanish Psychiatric Concept Normalization

Integración del razonamiento de LLM y UMLS para la normalización de conceptos psiquiátricos en español

Sergio Rubio-Martín[1], Arturo Crespo-Álvaro[1], María Teresa García-Ordás[1], Antonio Serrano-García[2], Clara Margarita Franch-Pato[2], Alicia Merayo-Corcoba[1], José Alberto Benítez-Andrades[1]

[1]ALBA Research Group, Universidad de León
[2]Servicio de Psiquiatría, Complejo Asistencial Universitario de León (CAULE)

{srubm, acrea, mgaro, amerc, jbena}@unileon.es,
drantonioserranogarcia@gmail.com,
clarafranch@hotmail.com

Abstract: This paper addresses the normalization of clinical concepts in Spanish psychiatric emergency notes, a particularly challenging task due to the use of informal, ambiguous, and context-dependent language, where lexical similarity alone is insufficient to ensure reliable mapping to standardized terminologies. We propose a hybrid architecture that combines semantic retrieval over a UMLS-based knowledge graph with a context-aware reasoning layer guided by clinical entity categories, implemented using LLMs. The method operates on entities extracted through a transformer-based NER model and adapts normalization to six clinical categories. Experiments were conducted on a clinically validated ground truth of 768 entities from 100 psychiatric emergency notes collected at CAULE. Results show substantial improvements over retrieval-only baselines, reaching 70.0% Hit@3 and 73.6% Hit@10 with MiniLM, and 60.9% Hit@3 and 63.3% Hit@10 with Multilingual BERT. Error analysis indicates strong performance in semantically complex categories, while pharmacological entities remain challenging. These findings highlight the effectiveness of combining structured knowledge with contextual reasoning for clinical concept normalization.
Keywords: Concept Normalization, Clinical Text, Psychiatry, LLM.

Resumen: Este artículo aborda la normalización de conceptos clínicos en notas psiquiátricas de urgencias en español, una tarea particularmente compleja debido al uso de lenguaje informal, ambiguo y dependiente del contexto, donde la similitud léxica por sí sola resulta insuficiente para garantizar un mapeo fiable a terminologías estandarizadas. Proponemos una arquitectura híbrida que combina recuperación semántica sobre un grafo de conocimiento basado en UMLS con una capa de razonamiento consciente del contexto, guiada por categorías de entidades clínicas, implementada mediante LLMs. El método opera sobre entidades extraídas mediante un modelo NER basado en transformadores y adapta la normalización a seis categorías clínicas. Los experimentos se realizaron sobre un ground truth validado clínicamente compuesto por 768 entidades de 100 notas psiquiátricas de urgencias recogidas en el CAULE. Los resultados muestran mejoras sustanciales respecto a los métodos basados únicamente en recuperación, alcanzando un Hit@3 del 70,0% y un Hit@10 del 73,6% con MiniLM, y un Hit@3 del 60,9% y un Hit@10 del 63,3% con Multilingual BERT. El análisis de errores indica un buen desempeño en categorías semánticamente complejas, aunque las entidades farmacológicas siguen siendo las más difíciles de normalizar. Estos resultados destacan la eficacia de combinar conocimiento estructurado con razonamiento contextual para la normalización de conceptos clínicos.
Palabras clave: Normalización de Conceptos, Textos Clínicos, Psiquiatría, LLM.

1 Introduction

Mental health has become a central concern in modern healthcare systems due to its growing prevalence and long-term impact on individuals and society (World Health Organization, 2025). Psychiatric conditions such as depression, anxiety, and substance-related disorders affect a significant portion of the population and often require sustained clinical attention. In recent years, an increase in both incidence and severity has been observed, particularly among younger populations. In Spain, this trend is reflected in the notable rise of psychiatric hospitalizations among teenagers over the last two decades (Soriano et al., 2025). These developments not only highlight the scale of the problem, but also expose limitations in how mental health information is currently managed and utilized.

Mental disorders rarely occur in isolation. Their interaction with physical conditions has been widely documented, revealing complex comorbidity patterns that influence diagnosis, treatment, and prognosis (Hanna et al., 2024; Martin Schulze, 2025). Consequently, accessing structured and semantically consistent clinical information is essential for enabling integrated and effective healthcare. However, a large portion of psychiatric knowledge remains embedded in narrative clinical records, limiting its direct usability in computational systems.

Electronic Health Records (EHRs) contain a wealth of psychiatric information, particularly within free-text clinical notes. These narratives are often written under time constraints and reflect the subjective interpretation of clinicians, resulting in heterogeneous, informal, and sometimes ambiguous language (Wang et al., 2023). While rich in detail, this variability hinders their integration into data-driven pipelines, reducing their potential for secondary analysis, clinical decision support, and interoperability.

Natural Language Processing (NLP) techniques have made significant progress in extracting structured information from such texts. In particular, Named Entity Recognition (NER) enables the identification of clinically relevant mentions such as symptoms, diagnoses, and treatments. Nevertheless, in the psychiatric domain, extraction alone does not guarantee meaningful structuring. Many expressions found in clinical narratives do not directly correspond to standardized medical terminology, as they may be indirect, context-dependent, or linguistically imprecise.

This issue is especially evident in psychiatry, where clinicians frequently describe patient states using non-literal or interpretative language. Expressions referring to behavioral patterns, perceptual disturbances, or emotional states often require clinical inference to be properly understood. As a result, mapping these mentions to standardized concepts cannot rely only on lexical similarity, but demands a deeper level of semantic interpretation. Concept normalization addresses this challenge by linking textual mentions to controlled vocabularies such as the Unified Medical Language System (UMLS), where each concept is represented by a unique identifier. This process is essential for transforming narrative descriptions into structured interoperable data that can be leveraged in downstream applications. However, in Spanish psychiatric contexts, normalization remains particularly difficult due to limited lexical coverage, variability in expression, and the presence of informal or domain-specific language.

In our previous work presented at CIABiomed, we explored the use of candidate retrieval strategies for concept normalization, combining embedding-based similarity and lexical matching approaches. Although these methods provided a baseline for linking entities to UMLS concepts, the results revealed important limitations when dealing with ambiguous and context-sensitive expressions.

This work advances clinical concept normalization by introducing a semantic interpretation module based on Large Language Models (LLM), designed to perform context-aware reasoning over extracted entities. Unlike retrieval-only approaches, this module enables the system to refine entity representations, resolve ambiguities, and select the most appropriate concepts based on contextual and clinical semantics rather than surface similarity alone. Furthermore, the pipeline follows an entity-aware design, where the normalization process is guided by clinically meaningful entity categories, allowing the model to adapt its reasoning strategy to the specific nature of each mention. Embedding-based retrieval is used as a supporting mechanism to provide candidate concepts, which are subsequently disambiguated through the proposed interpretation module.

The proposed approach is evaluated on a dataset of Spanish psychiatric clinical notes, demonstrating substantial improvements over retrieval-based strategies. By combining structured medical knowledge with contextual reasoning, this work moves toward more robust and semantically grounded representations of psychiatric information.

2 Related Work

Early approaches to clinical concept normalization relied on rule-based systems and dictionary matching techniques, which achieved acceptable precision in controlled domains but showed limited generalization when facing lexical variability or multilingual scenarios. These limitations motivated the transition toward more robust methods based on semantic representations.

In recent years, research has increasingly focused on embedding-based approaches, changing from surface-level matching to contextual semantic understanding. Models such as BioBERT and multilingual variants of BERT enable the representation of both clinical mentions and ontology concepts within shared vector spaces, facilitating candidate retrieval beyond exact lexical similarity (Zeakis et al., 2023; Hier, Do, y Obafemi-Ajayi, 2025). Additionally, incorporating ontological structure into the normalization process has been shown to improve disambiguation performance, particularly in cases involving ambiguous or polysemous terms (Wulff y Mata, 2025). More recent studies have extended this line of work by explicitly integrating relational and hierarchical knowledge into representation learning, showing that concept-aware embedding spaces can improve candidate ranking and semantic consistency during normalization (Singh, Krishnamoorthy, y Ortega, 2024; Callahan et al., 2023).

Parallel developments, such as multilingual approaches, have extended normalization techniques to languages beyond English, including Spanish, French, Serbian, and Chinese. However, the scarcity of annotated clinical corpora remains a major challenge in these settings (Zahra y Kate, 2024). To overcome this data scarcity, new resources like the RuCCoN dataset have been introduced, providing manually annotated clinical texts linked to UMLS concepts and contributing to the expansion of multilingual normalization capabilities (Nesterov et al., 2025). Recent shared-task and corpus-based studies have also shown that multilingual normalization can be approached through cross-lingual representation spaces, terminology translation, and candidate reranking strategies, although performance remains sensitive to domain specificity and knowledge base coverage (Gallego y Veredas, 2024; Vassileva et al., 2024).

Recent work has also explored few-shot and zero-shot learning strategies, particularly in domains with limited labeled data. Models such as BioLORD incorporate ontology-aware training mechanisms to enhance generalization and enable concept normalization in low-resource scenarios (Remy, Scaboro, y Portelli, 2023). Furthermore, LLMs have begun to play a relevant role in this context, showing the ability to support normalization through generative or reasoning-based approaches, especially when guided by structured prompts or external knowledge sources (Park et al., 2026; Zhang et al., 2026). At the same time, the integration of ontologies into semantic inference processes has gained attention, suggesting a convergence between concept normalization and automated clinical reasoning (Conceição et al., 2023; Abdel-salam, Adewunmi, y Akinwale, 2024). These approaches emphasize that normalization should not be treated as an isolated task, but rather as part of a broader framework for semantic understanding.

Despite these advances, most existing work has focused on general biomedical domains and predominantly on English-language data. In contrast, psychiatric clinical narratives present unique challenges due to their inherent subjectivity, linguistic variability, and frequent use of indirect or interpretative expressions. While some studies have demonstrated the ability of embedding-based representations to capture complex patterns in psychiatric data, such as identifying subgroups within schizophrenia (Brodersen et al., 2014) or modeling mental disorders as dynamic symptom networks (Gauld, 2022), these efforts do not directly address the problem of concept normalization.

In addition, recent work has shown that deep learning models are capable of capturing clinically meaningful patterns in psychiatric narratives, enabling the discrimination between diagnoses with overlapping symptomatology. For instance, it has been demonstrated that models can successfully classify clinical notes associated with disorders such as adjustment disorder and anxiety, even when these conditions share similar symptom profiles (Rubio-Martín et al., 2025b). These findings highlight the potential of deep learning to extract and represent latent clinical knowledge from psychiatric text, supporting downstream tasks like classification and prediction. However, this capability does not directly translate into effective concept normalization. While models can learn discriminative representations, mapping these representations to standardized terminologies such as UMLS remains challenging, particularly when dealing with implicit, colloquial, or context-dependent expressions. Recent studies focusing on clinical narratives in Spanish have further confirmed that lexical variability, multilingual transfer, and terminology coverage substantially affect normalization performance, even when using transformer-based architectures and hybrid retrieval strategies. Moreover, evaluations on Spanish clinical corpora show that normalization performance still depends heavily on the quality of candidate generation, the underlying terminology resource, and the ability to resolve ambiguity beyond surface similarity (Báez et al., 2025).

As a result, the normalization of psychiatric entities, particularly in Spanish, remains an open problem. The presence of colloquial language, implicit clinical meaning, and high semantic ambiguity complicates the alignment with standardized terminologies like UMLS for example. Although embedding-based and hybrid retrieval approaches have improved candidate generation, they often fail to capture the contextual interpretation required for accurate disambiguation.

In this context, our work builds upon previous advances by proposing a hybrid framework that combines semantic retrieval with context-aware reasoning guided by LLMs. This approach aims to bridge the gap between narrative psychiatric language and structured clinical knowledge, improving the reliability of concept normalization in this challenging domain.

3 Materials and Methods

3.1 Dataset and Ground Truth

This study is based on a corpus of Spanish-language psychiatric clinical notes collected from the Psychiatry Service of the Complejo Asistencial Universitario de León (CAULE). The dataset comprises records generated in emergency care settings, corresponding to unscheduled consultations rather than routine follow-up visits. As such, the notes capture the acute, dynamic, and often incomplete nature of real-world psychiatric assessment, where clinical information is recorded under time constraints and evolving conditions. All documents were fully anonymized in accordance with the General Data Protection Regulation (GDPR), and their use for research purposes was approved by the regional research ethics committee (approval ID: 2303). The complete dataset consists of 5,332 free-text clinical reports collected between 2017 and 2022, authored by multiple clinicians operating in high-pressure emergency environments. Consequently, the notes exhibit substantial heterogeneity in linguistic style and structure, including informal expressions, domain-specific abbreviations, shorthand notations, and occasional orthographic inconsistencies. Additionally, the same clinical concept is frequently expressed using different linguistic forms, introducing a high degree of semantic variability.

From this corpus, a subset of 100 clinical notes was selected for this study. The selection was not random but guided by variability criteria, aiming to preserve diversity in narrative style, clinical content, and entity distribution while maintaining a manageable size for expert validation. A total of 768 clinical entities were considered in this work. These entities belong to six predefined categories commonly found in psychiatric clinical documentation: Motive of Consultation (MC), Personal Clinical History (APC), Personal Pharmacological History (APF), Diagnosis (DX), Pharmacological Treatment (TTF) and Treatment of Action (TTA). On average, each note contains between 7 and 8 relevant entities, providing a balanced representation across categories.

All entities were previously extracted and subsequently reviewed by a domain expert in psychiatry. The same expert manually linked each entity to its corresponding concept in the Unified Medical Language System (UMLS), resulting in a clinically validated ground truth. This gold standard serves as the reference for evaluating the normalization strategies proposed in this work.

3.2 Entity Extraction (NER Stage)

The entities used in this study were obtained through a NER process developed in a previous stage of this research (Rubio-Martín et al., 2025a). Specifically, we employed a fine-tuned transformer-based model for Spanish clinical text, which demonstrated the best performance among the evaluated configurations. This model was trained to identify and classify entities according to the six categories defined above.

Rather than focusing on the extraction task itself, this work assumes the NER stage as a completed step and uses its output as input for the normalization pipeline. This design reflects a realistic scenario in which entity extraction and normalization are treated as sequential but distinct components within a clinical NLP system. Table 1 summarizes the distribution of entities across the different categories considered in this study.

Label	Count	Avg. per note
MC	119	1.19
APC	133	1.33
APF	171	1.71
DX	111	1.11
TTF	149	1.49
TTA	85	0.85

Table 1: Distribution of extracted entities by category.

3.3 From Retrieval-Based Normalization to Hybrid Reasoning

In our previous work, we addressed the problem of clinical concept normalization through a set of candidate retrieval strategies based on both semantic and lexical similarity. Specifically, we evaluated embedding-based approaches using MiniLM and Multilingual BERT, as well as rule-based methods such as recursive lexical matching and direct queries to the UMLS API (Rubio-Martín et al., 2026). These methods aimed to retrieve a ranked list of candidate concepts for each extracted entity.

Although these approaches provided an initial solution to the normalization task, the results revealed important limitations. Embedding-based methods showed low performance, achieving Hit@3 values of 0.16 and 0.18 for MiniLM and Multilingual BERT, respectively. Lexical matching strategies, while computationally efficient, were highly dependent on surface form similarity and failed to generalize to more complex expressions. The best-performing approach, based on UMLS API queries, reached a Hit@3 of 0.56, demonstrating the importance of leveraging external knowledge sources. However, even in this case, performance remained insufficient for reliable normalization in psychiatric narratives.

These limitations can be attributed to the intrinsic characteristics of psychiatric clinical language. Unlike more structured biomedical domains, psychiatric notes frequently contain indirect, context-dependent, or colloquial expressions that do not have a direct lexical correspondence with standardized terminology. As a result, approaches based only on similarity, either lexical or semantic, are unable to capture the intended clinical meaning in many cases. This observation highlights the need for a more flexible normalization strategy capable of incorporating contextual interpretation into the decision making process. In particular, it suggests that the normalization task should not be treated exclusively as a retrieval problem, but rather as a combination of candidate generation and semantic reasoning.

Building upon these observations, we propose a hybrid normalization architecture that integrates embedding-based candidate retrieval with a reasoning layer guided by LLMs. This reasoning layer is implemented using a modular orchestration framework that integrates LLMs and external knowledge sources within a structured processing pipeline. The core contribution lies in the integration of semantic retrieval, structured medical knowledge, and context-aware reasoning. In this context, LangChain is not used as a standalone solution, but as an intermediate component that facilitates the selection and validation of candidate concepts.

Unlike traditional approaches based on candidate reranking, the proposed method follows a structured three-stage process that integrates semantic reinterpretation, candi-

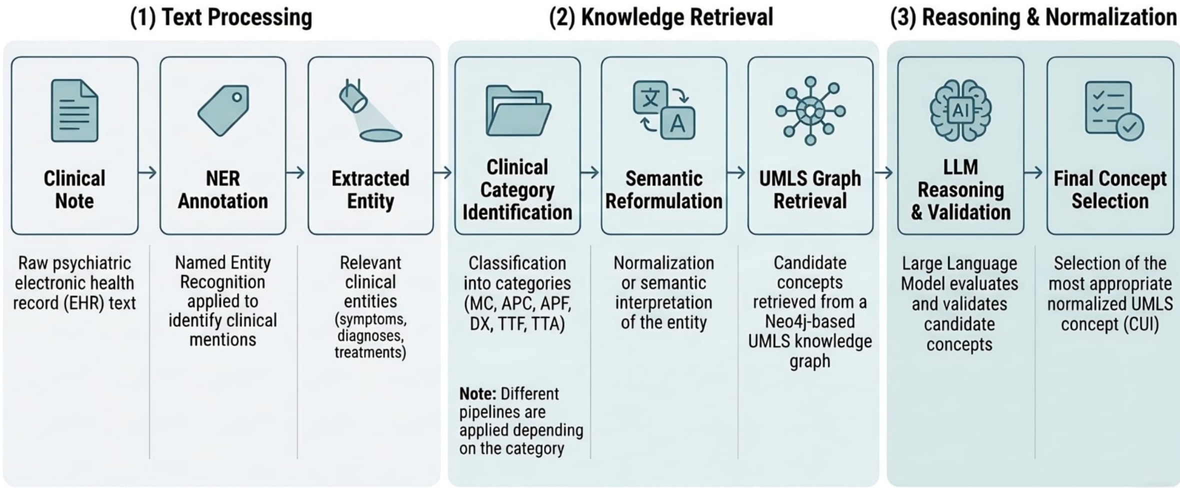

Figure 1: Full methodological pipeline followed.

date retrieval, and context-aware validation. First, each extracted entity is analyzed and processed according to its category, including reformulation when necessary to more accurately represent its underlying clinical semantics. Second, a set of candidate concepts is retrieved via embedding-based similarity search over the knowledge graph. Finally, a reasoning layer assesses the candidates within their contextual setting and selects the most appropriate UMLS concept, prioritizing semantic coherence over surface-level similarity.

To support the candidate retrieval stage, we constructed a structured semantic environment based on a partial extraction of the UMLS Metathesaurus, restricted to concepts relevant to psychiatry and psychology, including diseases, symptoms, substances, and pharmacological treatments. This resulted in a knowledge graph comprising 12,429 nodes, each of them representing a clinical concept identified by its corresponding identifier (CUI). The graph was implemented in Neo4j, enabling flexible and efficient access to semantic relationships. Each node was enriched with relevant attributes, including semantic type, associated synonyms, and multilingual labels in both English and Spanish, facilitating alignment between clinical notes and the predominantly English-based structure of UMLS.

In addition, each concept was associated with vector representations generated using MiniLM and Multilingual BERT. This allowed the system to perform both lexical and semantic retrieval, extending candidate generation beyond exact string matching. By combining structured ontological information with embedding-based representations, the knowledge graph provides a robust foundation for the normalization process.

A central contribution of the proposed approach is its entity-aware design, where the normalization strategy is explicitly adapted to each clinical category, enabling more accurate interpretation of context-dependent and semantically ambiguous expressions. Each category of entity (e.g., MC, APC, APF, DX, TTF, TTA) presents distinct linguistic and semantic characteristics, which require different interpretation mechanisms. For this reason, the reasoning process is adapted to each category through tailored prompts that guide the model toward more accurate and context-sensitive normalization.

This hybrid approach enables a tighter integration between structured medical knowledge and contextual language understanding, addressing the limitations observed in retrieval-based methods and improving the robustness of concept normalization in psychiatric clinical narratives. The full pipeline is shown in Figure 1.

3.4 Hardware and Software Setup

The computational infrastructure consisted of an Intel Core i7-13700 processor, 32 GB of RAM, and an NVIDIA RTX 4060 Ti GPU. The experimental pipeline was developed in Python 3.10 within a local execution environment. The implementation integrates several libraries commonly used in clinical Natural Language Processing workflows. Transformer-based models were managed using the HuggingFace Transformers framework, while data preprocessing and handling were performed using Pandas and the HuggingFace Datasets library. The semantic retrieval component relied on embeddings generated through the SentenceTransformers library, enabling the representation of both entities and candidate concepts in a shared vector space. Similarity between vectors was computed using cosine similarity.

For knowledge representation and querying, a Neo4j graph database was deployed locally to store the UMLS-derived concept graph. Manual annotation and validation of the clinical entities were carried out using the academic version of Label Studio, which provides support for span-based annotation in textual data (Tkachenko et al., 2020-2025).

4 *Experiments and Results*

4.1 Reasoning-Enhanced Clinical Concept Normalization

The results obtained in the previous stage revealed clear limitations of normalization approaches based exclusively on lexical matching or embedding similarity. While these methods were effective in retrieving candidate concepts, they struggled to correctly interpret ambiguous, colloquial, and context-dependent expressions, which are particularly common in psychiatric clinical narratives.

To address this limitation, we propose a hybrid normalization architecture that integrates embedding-based candidate retrieval with a reasoning layer guided by LLMs. Our contribution lies in a category-aware normalization strategy that combines semantic candidate retrieval from a psychiatry-oriented UMLS graph with contextual reasoning for final concept selection. The reasoning layer is implemented through a structured reasoning pipeline that enables the orchestration of language models and external knowledge sources within the normalization process (Yang et al., 2025). In our implementation, LangChain is used as an intermediate component that facilitates the selection and validation of candidate concepts, rather than as the core contribution of the method.

The proposed approach integrates semantic retrieval with contextual reasoning. Given an extracted entity, the system first performs a semantic reinterpretation step, refining the expression when necessary to better reflect its underlying clinical meaning. Subsequently, candidate concepts are retrieved from the knowledge graph, and finally, the reasoning layer evaluates these candidates in context to select the most appropriate UMLS concept.

The reasoning component was implemented using the GPT-4o model deployed via the Azure OpenAI service, enabling controlled and reproducible inference within the pipeline. This setup allows the system to incorporate contextual understanding while maintaining integration with structured knowledge sources such as UMLS.

The proposed reasoning framework enhances the normalization process by introducing context-aware methodological improvements. First, it enables seamless integration between language models and external knowledge bases. Second, it provides traceability of the decision making process, including intermediate reasoning steps and candidate evaluation. Third, it allows dynamic adaptation of prompts depending on the type of entity being processed.

Overall, this approach transforms the normalization task from a purely retrieval-based problem into a hybrid process that combines candidate generation with context-aware semantic reasoning, which is particularly beneficial in the psychiatric domain.

4.2 Entity-Aware Normalization Strategies

The proposed framework was implemented using a three-role prompting strategy combined with few-shot learning to improve performance. This stage begins with the system role, which provides the model with task-specific context and instructions. Subsequently, the user role introduces a set of representative examples, while the assistant role supplies the corresponding expected outputs. Finally, the user role presents the target instances to be solved by the model. The previously described framework is adapted to the six entity categories defined during the extraction stage, namely MC, APC, APF, DX, TTF, and TTA. Each category exhibits distinct linguistic and semantic properties, thereby requiring tailored reasoning strategies to ensure accurate normalization.

For Motive of Consultation (MC), entities are often expressed using informal or indirect language. To facilitate interpretation, three semantic subgroups were defined: (i) direct clinical expressions (e.g., "anxiety", "persistent sadness"), (ii) abnormal behaviors described narratively (e.g., "metió el pie en una poza"), and (iii) perceptual alterations associated with hallucinations (e.g., "ve cosas", "escucha voces").

In the case of Personal Clinical History (APC), normalization focused on distinguishing between different clinical contexts, including (i)diseases, (ii)medical procedures, (iii)self-harm ideation, (iv)non-lethal self-harm behaviors, and (v)suicide attempts. This categorization allows a more precise mapping of historical clinical information to standardized concepts.

For Personal Pharmacological History (APF), entities were classified into (i)pharmacological treatments and (ii)substance use. This distinction is particularly relevant due to the variability in how medications and substances are referenced in clinical text.

Diagnosis (DX) entities were primarily mapped to psychiatric disorders, with special consideration for cases involving self-harm-related conditions. These entities tend to be more aligned with standardized terminology, facilitating normalization. When applicable, the system also assigns ICD-10 codes for disease-related entities.

Pharmacological Treatment (TTF) required distinguishing between (i)active prescriptions and (i)discontinued medications. The normalization process also accounts for dosage and administration details when explicitly mentioned in the entity.

Finally, Treatment Action (TTA) represents structured clinical decisions, such as (i)discharge, (ii)admission, or (iii)follow-up. Due to their limited variability, these entities present lower ambiguity and are more consistently mapped to standardized concepts.

Across all categories, the reasoning layer performs entity refinement, semantic interpretation, and candidate validation. Additionally, pharmacological entities are normalized to their generic names when commercial names are detected.

This entity-aware design enables the system to adapt to the specific characteristics of each category, improving the consistency and accuracy of the normalization process while maintaining traceability of the decisions.

4.3 Results

To evaluate the proposed normalization strategy, a dataset of 768 clinical entities was used. Unlike the baseline approaches, the proposed method first reinterprets and refines each entity according to its clinical category, then retrieves candidate concepts from the UMLS-based knowledge graph, and finally validates the most appropriate mapping based on semantic coherence.

The evaluation focuses on embedding-based retrieval methods, comparing their performance when integrated into the proposed normalization strategy. Results are reported using Hit@k metrics with $k \in \{3, 5, 7, 10\}$, which measure whether the correct concept appears within the top-k retrieved candidates. Table 2 summarizes the performance obtained with MiniLM and Multilingual BERT when candidate retrieval is combined with entity reinterpretation and semantic validation. In both cases, the results show a substantial improvement over retrieval-only approaches, confirming that candidate generation alone is insufficient when clinical mentions require contextual interpretation. For visualization purposes, Figure 2 presents the corresponding performance trends across different values of k.

Both models exhibit a consistent increase in performance as the number of considered candidates grows. MiniLM achieves the best overall results, reaching a Hit@3 of 70% and exceeding 73% at Hit@10. In contrast, Multilingual BERT shows lower results but still competitive performance, ranging between 61% and 63%. These results suggest that the reasoning layer benefits both models, although its impact is more pronounced when combined with MiniLM embeddings.

To provide a more detailed analysis, Table 3 reports the number of errors per entity category. The results reveal that APF and TTF remain the most challenging categories, particularly for Multilingual BERT. This can be attributed to the high variability of pharmacological expressions, including differences in dosage, formulation, and the use of commercial names.

In contrast, TTA exhibits the lowest error rates, reflecting the more structured and limited nature of these entities (e.g., discharge, admission, follow-up), which facilitates consistent mapping to standardized concepts.

Overall, the results show that normaliza-

Model	Hit@3	Hit@5	Hit@7	Hit@10
MiniLM	534 (70.0%)	546 (71.1%)	550 (71.6%)	565 (73.6%)
Multilingual BERT	468 (60.9%)	470 (61.2%)	475 (61.8%)	486 (63.3%)

Table 2: Normalization performance with the proposed reasoning-enhanced approach.

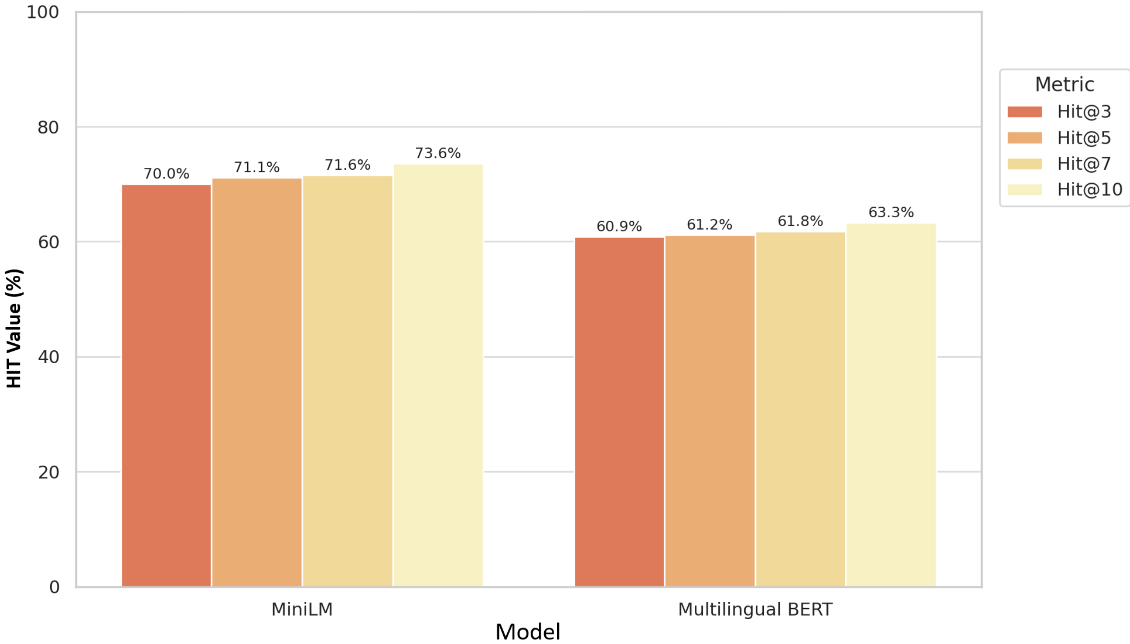

Figure 2: Model performance across different Hit@k values.

Category	MiniLM	MultiL. BERT
MC	45	46
APC	29	34
APF	**52**	**83**
DX	37	46
TTF	**59**	**79**
TTA	12	12

Table 3: Number of normalization errors by entity category.

tion performance improves when the system combines three complementary steps: reinterpretation of the clinical entity, retrieval of semantically relevant candidates, and final validation based on contextual coherence. This combination is particularly useful in psychiatric clinical language, where ambiguity, indirect expression, and semantic variability often limit the effectiveness of retrieval-only approaches.

5 *Discussion and conclusions*

The comparative analysis between retrieval-based approaches and the proposed reasoning-enhanced architecture highlights the impact of incorporating contextual interpretation into the normalization process. In the baseline scenario, methods based only on lexical matching or embedding similarity achieved limited performance. In particular, MiniLM and Multilingual BERT obtained Hit@3 scores of 16% and 18%, respectively, reflecting the difficulty of mapping Spanish psychiatric expressions to standardized UMLS concepts when relying exclusively on surface-level similarity. Even the strongest baseline, based on UMLS API queries, reached a Hit@3 of 56% and a Hit@5 of 58%, indicating that terminological coverage alone is insufficient to handle semantically complex clinical language.

After integrating the reasoning layer, the results improved substantially. MiniLM achieved a Hit@3 of 70% and a Hit@5 of 71%, representing an improvement of more than 50 percentage points compared to the baseline. Similarly, Multilingual BERT increased its performance from 18% to 61% in Hit@3 and up to 63% in Hit@5. These results demonstrate that incorporating contextual reasoning enables the system to better interpret clinical meaning and select more appropriate candidate concepts, particularly in cases where lexical similarity is insufficient.

From a qualitative perspective, the proposed approach proved especially effective in resolving semantically ambiguous or indirect expressions. Categories such as MC and TTF benefited from the system's ability to reinterpret non-literal language (e.g., expressions referring to hallucinations or treatment changes), while structured categories such as TTA exhibited consistently low error rates due to their more constrained nature. This behavior confirms that the primary limitation of retrieval-based methods lies not in candidate generation, but in the lack of contextual interpretation.

Despite these improvements, several limitations must be acknowledged. First, the evaluation was conducted on a controlled subset of 100 clinical notes, which, although representative, may limit the generalizability of the results. Second, the use of LLMs introduces additional computational costs and dependency on external services, which may affect scalability in real-world clinical environments. Furthermore, some cases of semantic overgeneralization were observed, where the model selected broader concepts than expected (e.g., mapping "brote psicótico" to more generic psychotic disorders), reflecting the inherent flexibility of generative models.

In summary, this work shows that integrating a reasoning layer within the normalization pipeline helps to improve both the accuracy and semantic coherence of concept mapping in psychiatric clinical narratives. The proposed approach effectively bridges the gap between informal clinical language and structured medical knowledge, enabling more reliable normalization in complex linguistic contexts.

Future work will focus on extending this framework in several directions. First, scaling the evaluation to larger and more diverse datasets will be essential to validate the robustness of the approach. Second, incorporating more advanced prompting strategies or fine-tuned language models could further improve performance in challenging cases. Finally, deeper integration with the knowledge graph, including iterative enrichment and feedback loops, may enhance both normalization accuracy and downstream clinical applications.

Acknowledgement

This work was supported by Grant PID2023-146168OA-I00 "NLP-Driven Insight Engine for Suicide Attempt Detection in EHR (SUICIDETECT)" funded by MICIU/AEI/10.13039/501100011033 and by the "European Union".

Bibliografía

Abdel-salam, R., M. Adewunmi, y M. Akinwale. 2024. CaresAI at SemEval-2024 task 2: Improving natural language inference in clinical trial data using model ensemble and data explanation. En *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, páginas 1905–1911. Association for Computational Linguistics, Junio.

Brodersen, K. H., L. Deserno, F. Schlagenhauf, Z. Lin, W. D. Penny, J. M. Buhmann, y K. E. Stephan. 2014. Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage: Clinical*, 4:98–111.

Báez, P., L. Campillos-Llanos, F. Núñez, y J. Dunstan. 2025. Entity normalization in a spanish medical corpus using a umls-based lexicon: findings and limitations. *Language Resources and Evaluation*, 59(2):1013–1041.

Callahan, T. J., A. L. Stefanski, J. M. Wyrwa, C. Zeng, A. Ostropolets, J. M. Banda, W. A. Baumgartner, R. D. Boyce, E. Casiraghi, B. D. Coleman, J. H. Collins, S. J. D. Davies, J. A. Feinstein, A. Y. Lin, B. Martin, N. A. Matentzoglu, D. Meeker, J. Reese, J. Sinclair, S. B. Taneja, K. E. Trinkley, N. A. Vasilevsky, A. E. Williams, X. A. Zhang, J. C. Denny, P. B. Ryan, G. Hripcsak, T. D. Bennett, M. A. Haendel, P. N. Robinson, L. E. Hunter, y M. G. Kahn. 2023. Ontologizing health systems data at scale: making translational discovery a reality. *npj Digital Medicine*, 6(1):89.

Conceição, S. I. R., D. F. Sousa, P. Silvestre, y F. M. Couto. 2023. lasigeBioTM at SemEval-2023 task 7: Improving natural language inference baseline systems with domain ontologies. En *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, páginas 10–

15. Association for Computational Linguistics, Julio.

Gallego, F. y F. J. Veredas. 2024. Recognition and normalization of multilingual symptom entities using in-domain-adapted bert models and classification layers. *Database*, 2024:baae087, 08.

Gauld, C. 2022. From psychiatric kinds to harmful symptoms. *Synthese*, 200(6):440, Octubre.

Hanna, M. R., A. Caspi, R. M. Houts, T. E. Moffitt, y F. A. Torvik. 2024. Co-occurrence between mental disorders and physical diseases: a study of nationwide primary-care medical records. *Psychological Medicine*, 54(15):4274–4286.

Hier, D. B., T. S. Do, y T. Obafemi-Ajayi. 2025. A simplified retriever to improve accuracy of phenotype normalizations by large language models. *Frontiers in Digital Health*, Volume 7 - 2025.

Martin Schulze, Tim Reimers, J. H. T. B. S. B. A. G. S. S. 2025. Management of somatic comorbidities in psychiatric emergencies: insights from a real-world cohort study. *Psychiatry Research*, página 116630.

Nesterov, A., A. Sakhovskiy, I. Sviridov, A. Valiev, V. Makharev, P. Anokhin, G. Zubkova, y E. Tutubalina. 2025. Ruccod: Towards automated icd coding in russian. *ArXiv*, abs/2502.21263.

Park, Y., H. Kang, J. Kim, S.-Y. Shin, D. Cho, S. Y. Rhee, H. S. Park, K.-J. Lee, y S. Bae. 2026. Development and evaluation of snomed ct automated mapping tool: Advancing terminology standardization and semantic interoperability. *JMIR Medical Informatics*, 14:e82670, Mar.

Remy, F., S. Scaboro, y B. Portelli. 2023. Boosting adverse drug event normalization on social media: General-purpose model initialization and biomedical semantic text similarity benefit zero-shot linking in informal contexts. *ArXiv*, abs/2308.00157.

Rubio-Martín, S., A. Crespo-Álvaro, M. T. García-Ordás, A. Serrano-García, C. M. Franch-Pato, y J. A. Benítez-Andrades. 2026. Concept normalization in psychiatry: Comparing embedding and lexical methods for spanish clinical text. En A. López Fernández A. Rodríguez-González R. Leirós-Rodríguez C. Mata Miquel, y V. M. González Suárez, editores, *Artificial Intelligence in Biomedicine*, páginas 282–296, Cham. Springer Nature Switzerland.

Rubio-Martín, S., A. Crespo-Álvaro, M. T. García-Ordás, A. Serrano-García, C. M. Franch-Pato, y J. A. Benítez-Andrades. 2025a. Fine-tuning transformer models for structuring spanish psychiatric clinical notes. En *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, páginas 97–102.

Rubio-Martín, S., M. T. García-Ordás, A. Serrano-García, C. M. Franch-Pato, A. Crespo-Álvaro, y J. A. Benítez-Andrades. 2025b. Classification of psychiatry clinical notes by diagnosis: a deep learning and machine learning approach. *PeerJ Computer Science*, 11:e3045.

Singh, A., S. Krishnamoorthy, y J. E. Ortega. 2024. Neighbert: Medical entity linking using relation-induced dense retrieval. *Journal of Healthcare Informatics Research*, 8(2):353–369.

Soriano, V., J. M. Ramos, M. I. López-Ibor, C. Chiclana-Actis, M. Faraco, J. González-Cabrera, E. González-Fraile, G. Mestre-Bach, H. Pinargote, M. Corpas, L. Gallego, O. Corral, y H. Blasco-Fontecilla. 2025. Hospital admissions in adolescents with mental disorders in Spain over the last two decades: a mental health crisis? *European Child & Adolescent Psychiatry*, 34(3):1125–1134, Marzo.

Tkachenko, M., M. Malyuk, A. Holmanyuk, y N. Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Vassileva, S., G. Grazhdanski, I. Koychev, y S. Boytcheva. 2024. Transformer-based approach for symptom recognition and multilingual linking. *Database*, 2024:baae090, 09.

Wang, W., D. Ferrari, G. Haddon-Hill, y V. Curcin, 2023. *Electronic Health Records as Source of Research Data*, páginas 331–354. Springer US, New York, NY.

World Health Organization. 2025. World mental health today: latest data.

Wulff, D. U. y R. Mata. 2025. Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*, 9:944 – 954.

Yang, J., L. Shu, H. Duan, y H. Li. 2025. Rdguru: A conversational intelligent agent for rare diseases. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6366–6378.

Zahra, F. A. y R. J. Kate. 2024. Obtaining clinical term embeddings from snomed ct ontology. *Journal of Biomedical Informatics*, 149:104560.

Zeakis, A., G. Papadakis, D. Skoutas, y M. Koubarakis. 2023. Pre-trained embeddings for entity resolution: An experimental analysis. *Proc. VLDB Endow.*, 16:2225–2238.

Zhang, G., Y. Fang, F. Chen, C. Ta, G. Hripcsak, P. Ryan, Y. Peng, y C. Weng. 2026. A neural-symbolic ai agent system for biomedical concept mapping. *npj Digital Medicine*.